

METHOD AND APPARATUS FOR EVALUATING MOLECULAR SIMILARITY

ins B1)

Field of the Invention

5 This invention relates to a method and apparatus for comparing molecules. It is especially useful in comparing individual molecules against a large library of molecules and has particular application in the pharmaceutical industry in drug development.

10

Background of the Invention

00644937-082300  
15 A molecule is normally thought of as a set of atoms of varying atomic type, with a certain bonding pattern. Indeed this "chemical" description can uniquely describe the molecule. It is the language that chemists use to compare and contrast different molecules. Efficient database models have been constructed to store such information for fast retrieval and storage. However, this form of description  
20 does not actually describe the three dimensional structure of the molecule, e.g. the positions of each atom, and since the interaction of molecules is a spatial event, the "chemical" description is incomplete for physical phenomena. One such  
25 phenomenon of commercial importance is the binding of drug molecules to sites of biological importance, such as the active areas or "sites" on protein surfaces, which is the mode of action of nearly all pharmaceuticals.

30 Drug molecules are often small, on the order of 20 atoms (excluding hydrogens). They interact with large

macromolecules such as proteins by binding to them. Through binding, the drug may activate or inhibit the normal action of the macromolecule. The binding occurs at specific sites on the macromolecule, and the basis of tight and specific  
5 binding is complementarity in shape and other properties, such as electrostatic, between the two molecules.

Pharmaceutical companies maintain computer databases of all molecules they have synthesized, plus other compounds  
10 available on the market. The use of these databases and the techniques of computer-aided drug design are beginning to replace trial and error lab testing in new drug development. Important components of this process are finding new small molecules similar in shape to ones known to bind a target,  
15 and designing new molecules to fit into known or hypothesized binding sites.

There have been many attempts to describe or "encode" the three dimensional information of molecules beyond a  
20 simple list of coordinates. Many involve the distances between pairs of atoms in a molecule, i.e. an atomistic approach akin to the chemical description but with extra, spatial degrees of freedom.

A more radical departure is to adopt an alternate  
25 representation of a molecule: the field representation. A field is essentially just a number assigned to every point in space. For instance, the air temperature in a room at every point in that room forms a field quantity. Molecules have  
30 one fundamental field associated with them, namely the quantum mechanical field that describes the probability of

electrons and nuclei existing at each point in space. However, this field can be thought of as giving rise to other, simpler fields of more use in understanding a molecule's properties. Chief amongst these are the steric and electrostatic fields, although others are used, such as the hydrophobic and the hydrogen bond potential field. An illustration of a Gaussian representation of a steric field is shown in Fig. 1. As is customary, on each contour line in the field, each point has equal value.

Steric fields describe the mass or shape of the molecule, and at the simplest level such a field might have a value of one inside the molecule and zero outside. Electrostatic fields represent the energy it would take to place an electron at a particular place in space, by convention positive if the energy is unfavorable and negative if favorable. These two fields are the most relevant for molecular interactions because of basic physical laws, i.e. that two molecules cannot overlap (steric repulsion) and that positive atoms like to be near negative atoms and vice versa. These are the basic components of molecular interactions.

If a molecule is known to bind and have effect on some biological target, it is of great commercial interest to find other molecules of similar shape and electrostatic properties (i.e. similar fields) since this enhances the likelihood of such molecules having similar biological activity. Since shape and electrostatic character are consequences of the underlying atoms, which can be efficiently encoded by a chemical description, such searches have traditionally been

002280" 4E6H960

performed at this level, by looking for molecules which are "chemically" similar. One disadvantage of this approach is that the relationship between chemical similarity and structural similarity is not precise; and chemically similar molecules may be structurally quite different. Another disadvantage is that a chemically similar compound may well be covered by the same patents as the original molecule. Finally, searching only chemically similar compounds inevitably means one will not find active molecules that are not chemically similar.

This latter point is key. David Weininger of Daylight Chemical Information Systems has reported an analysis that suggests there are  $10^{200}$  different molecules synthesisable by known means. (Only  $10^{107}$  molecules of typical drug size would fit in the known universe!). As such, any molecule, of any shape or electrostatic profile, has a potentially astronomical number of similarly shaped and charged "doppelgangers". Only a fraction of them are necessarily chemically similar. Hence by restricting the search to chemical similarity, a vast number of potential drug leads are never discovered.

Although  $10^{200}$  molecules is too large a number to ever enumerate, I believe that it is possible to determine bounds to the possible variations of molecular fields of this hypothetical set. Furthermore, I plan to compute for database storage a very large number of molecular structures (e.g. of the order of billions) that sample this range such that I am able to find a match, or "mimic", from this

collection to any novel structure presented. Such a database would be many thousands of times larger than any currently in existence and hence crucial to this plan is the efficient organization of such for fast search and retrieval of such  
5 mimics and the assessment of whether I have indeed "covered" chemical space. It is these problems that the present invention addresses.

### Prior Art

10 Much has been done in the use of molecular fields to compare and contrast molecules and to predict activity from such operations. Some of these approaches are described below. I believe that the crucial aspect of my approach  
15 which differs from all prior work is in the application of a particular property of field comparison, namely the "metric" property, and in a novel way to decompose fields into separable domains, wherein each is quantifiably similar to a geometrically simpler field.

20 The most widely known "field analysis" approach is that known as Comparative Molecular Field Activity (COMFA). See U.S. Patents 5,025,388 and 5,307,287 assigned to Tripos Inc. of St. Louis, Missouri. The idea behind COMFA is to take a  
25 series of molecules of known activity and to find which parts of these molecules are responsible for activity. The procedure is to first overlay the set of molecules onto each other such that the combined difference of the steric and electrostatic fields between all pairs of molecules is at a  
30 minimum. (The concept of overlaying, i.e. finding an



procedure then repeated until the number of compounds chosen reaches a desired threshold.

In both COMFA and the Chapman approach, field similarity is used as a tool to solve the alignment issue, and  
5 similarities or differences are then calculated. The value of the field similarity or difference is of secondary importance, it merely solves what is called the "assignment" problem, i.e. which atoms, or areas of a molecule's field are  
10 "equivalent".

In contrast, in Mestres et al., "a Molecular Field-Based Approach to Pharmacophoric Pattern Recognition," J. Molecular Graphics and Modelling, Vol. 15, pp. 114-121 (April 1997),  
15 molecules are aligned based upon the overlap of their steric or electrostatic fields, or by a weighted sum of the two. A similarity measure is defined that equals one when the fields are the same, and minus one when they are maximally different. The Mestres et al. work is embodied in a program  
20 called MIMIC, which performs global and local optimization of the field overlap. They note that there are several possible overlays that have the appearance of being the best overlap. These are so called "local minima", because while small displacements lead to a decrease in their similarity  
25 function, they may not be the best "global" solution. This is as expected since field overlay belongs to the class of problems known to have "multiple minima". Mathematically this is usually an intractable problem, solvable only by much  
30 computation, e.g., as is evident in the descriptions by Mestres et al. In fact, the multiple overlay solutions are

one of the key aspects of their work, in that one cannot be sure which is the most "biologically" relevant overlay, and what might be the correct weighting of steric to electrostatic fields.

10 An additional aspect considered by Mestres et al. is the issue of molecules existing in multiple structural conformations, i.e. energetically there may be more than one possible structure for a given molecule. Mestres et al.

15 calculate the similarity indexes of all pairs of conformations of a molecule and perform what is known as a principle component analysis (PCA). They do this to find representatives of all possible conformations that are most distinct. Although this procedure is really akin to finding

20 the dimensionality of the space in which these conformers exist, Mestres et al. do not use PCA for this purpose, but merely to cluster the conformers. They do not apply PCA to sets of different molecules, only to conformers of the same molecule, and they do not use any other "metric" property of their similarity measure. In fact they seem unaware of such.

There is an important distinction to be made between a "measure" of similarity and a "metric" of similarity, although these words are often used interchangeably. A

25 measure can be any quantity which has a correspondence with molecular similarity, i.e. the idea that the more similar the measure the more similar the compounds. A metric has a precise mathematical interpretation, namely that if the

30 metric, or more commonly the metric distance, between A and B is zero then the two items are the same item, that the



distance from A to B is the same as the distance from B to A, and that the distance from A to B plus the distance from B to a third compound C must be greater than the distance from A to C. This latter is called the "Triangle Inequality"

5 because the same conditions can be said of the sides of a triangle ABC. The Triangle Inequality, or metric upper bound, also leads to a lower bound, namely that in the case above, C can be no closer to A than the difference of these  
10 distances A to B and B to C.

In M. Petitjean, "Geometric Molecular Similarity from Volume-Based Distance Minimization-Application to Saxitoxin and Tetrodotoxin," J. Computational Chemistry, Vol. 16, No. 1, pp. 80-95 (1995), it is recognized that the quantity that  
15 measures the overlay of fields forms a metric quantity, and that the measure of the optimum overlay of two fields also forms a metric which is intrinsic to the molecule, i.e. independent of orientation or position.

*See 12*  
20 A metric distance may also be used in a technique called "embedding". The number of links between the elements of a set of N elements can be shown to be  $N*(N-1)/2$  and each link can be shown to be a metric distance. While a set of N elements has  $N*(N-1)/2$  distances, the set can always be  
25 represented by an ordered set of (N-1) numbers, i.e. I can "embed" from a set of distances to a set of N positions in (N-1) dimensional space. This is identical to Principle Component Analysis mentioned previously, except that with PCA  
30 one finds the most "important" dimensions, i.e. the "principal" directions, which carry most of the variation in

could be  
position. Typically with PCA one truncates the  
dimensionality at 2 or 3 for graphical display purposes. In  
general, the number of dimensions which reproduces the set of  
 $N*(N-1)/2$  distances within an acceptable tolerance may be  
5 much smaller than  $(N-1)$ , yet still be greater than 2 or 3.  
Hence one talks of "embedding into a hyper-dimensional  
subspace", where hyper-dimensional means more than 3  
dimensions, and subspace means less than  $(N-1)$ . Techniques  
10 for such an embedding are standard linear algebra. When  
applied to molecular fields, the result of embedding is a  
shape-space of  $M \leq N-1$  dimensions.

#### Summary of the Invention

15 I have invented several techniques for characterizing  
molecules based on the shapes of their fields. The minimal  
distance between two molecular fields is used as a shape-  
based metric, independent of the underlying chemical  
20 structure, and a high-dimensional shape space description of  
the molecules is generated. These attributes can be used in  
creating, characterizing, and searching databases of  
molecules based on field similarity. In particular, they  
25 allow searches of a database in sublinear time. The utility  
of this approach can be extended to automatically break  
molecules into a series of fragments by using an ellipsoidal  
Gaussian decomposition. Not only can these fragments then be  
analyzed by the shape metric technique described above, but  
30 the parameters of the decomposition themselves can also be

used to further organize and search databases. The ellipsoidal method can also be used to describe binding or active sites on macromolecules, providing a template for searching for complementary molecules in a database such as I  
5 describe. The most immediate application of these techniques is to pharmaceutical drug discovery and design.

In a preferred embodiment, I obtain the minimal distance between a first molecular field and a multiplicity N of other  
10 fields by: selecting a small number M of the fields, for each of the M fields determining its metric distance to all the other N fields, for each of the M fields, making an ordered list of the metric distances between that field and all the other N fields, determining the metric distances between the  
15 first field and each of the M fields, determining the metric distances between the first field and the fields on the ordered list associated with the M field that has the shortest metric distance between it and the first field.  
20 These metric distances are determined beginning with the field on the list that has the shortest metric distance between it and the M field and continuing such determination with fields having increasingly greater metric distances from the M field until a field is reached that has a metric  
25 distance from the M field that is more than twice the metric distance from the first field to the M field.

Advantageously, the invention is practiced on a computer, the determination of minimal distances and  
30 ellipsoidal Gaussian decomposition are implemented by computer programs and the molecular field information is

stored in a computer database. Illustrative apparatus for practicing the invention is a personal computer such as an IBM-compatible PC or a work station such as a Silicon Graphics Iris Indigo Elan 4000.

5

### Brief Description of the Drawings

These and other objects, features and advantages of my invention will be more readily apparent from the following  
10 detailed description of the invention in which:

Fig. 1 is an illustration of a Gaussian representation of a steric field;

Figs. 2A and 2B illustrate the results of overlaying two  
15 molecules using a prior art technique;

Fig. 3 is an illustration of a molecular field representation produced in accordance with the invention;

Figs. 4A, 4B and 4C are examples of three molecular field representations formed using increasing numbers of  
20 ellipsoidal Gaussian functions in accordance with the invention;

Fig. 5 is a flow chart illustrating one aspect of the invention;

Fig. 6 is a flow chart illustrating a second aspect of  
25 the invention; and

Fig. 7 is a flow chart illustrating a third aspect of the invention.

30

## Detailed Description of the Invention

### Definitions

The following defined terms are used in the detailed description:

5 A **structure**: A description of the three dimensional coordinates of each of the atoms that comprise a molecule. These coordinates may be found from experiment, e.g. X-ray crystallography, or by computer computation by one of many  
10 methods known in the field of molecular modeling.

A **conformer**: If a molecule has more than one structure, then each structure is referred to as a conformer of that  
15 molecule.

A **molecular field**, or **field**: A set of numbers that represent the value of some property in and around a molecule. Such  
20 numbers may be explicitly stated, or listed, for instance as values associated with each point on a regular lattice, or grid, which contains the molecule. Or they may be functionally implied. For instance, if a functional form for the field property is assigned to each atom then a mechanism  
25 exists to calculate the field value at any point in space. One example is the electrostatic potential around a molecule, one form of which may be calculated from the charge associated at each atom and the functional form for  
30 electrostatic potential from a single charge, i.e. Coulomb's Law.



$$F_M(\bar{r}) = \sum_{i=1}^N G_i(\bar{r}) \quad (2)$$

5

Exclusion Product Form:

$$F_M(\bar{r}) = 1 - \prod_{i=1}^N (1 - G_i(\bar{r})) \quad (3)$$

where  $F_M$  is the steric field of a molecule with structure and orientation M made up of N atoms represented by Gaussian  
15 functions  $G_i$  centered about each atom i. Each form has the property of having zero values away from the molecule and usually positive values "inside" the molecule. The exclusion product form expands to a sum of Gaussians, because the  
20 product of two Gaussians is itself a Gaussian function.

**Field Arithmetic:** Two fields are added together by adding their values at corresponding grid points, or by adding together the functional forms that define the field. A field can be "scaled" by multiplying either the value at each grid point, or the functional form, by a number. Two fields can be multiplied together by multiplying the values at corresponding grid points by each other or multiplying their functional forms together.

096472

5

10

20

30



if the molecule is represented on a grid as described above  
in Field Volume.

5 The **norm** of a field  $F_M$  is defined as the square root of the  
overlap of that field with itself, i.e.,

$$|F_M| = \sqrt{\int F_M(\bar{r}) F_M(\bar{r}) d\bar{r}} \quad (8)$$

10

The **norm of the field difference**, also referred to as the  
**field difference**  $D$ , is given by

15

$$D = |F_M - F_P| = \sqrt{\int (F_M(\bar{r}) - F_P(\bar{r}))^2 d\bar{r}} \quad (9)$$

20

The greater the field overlap, the less the field difference,  
as can be shown by rewriting the above equation as

$$25 \quad D^2 = \int F_M^2(\bar{r}) d\bar{r} + \int F_P^2(\bar{r}) d\bar{r} - 2 \int F_M(\bar{r}) F_P(\bar{r}) d\bar{r} \quad (10)$$

Only the third integral varies if  $F_M$  and  $F_P$  are moved relative  
30 to each other. But this third integral is equal to the

096527060

□

10

15

20

These are the upper and lower bounds, respectively, for the value of  $d_{AC}$ .

## 25

30





The Sum Form of the EGR is currently preferred.

5 The procedure for fitting one or more EGF's to a molecular  
field involves ascertaining the parameters of each EGF, i.e.,  
center (a,b,c), widths (u,v,w), and axes directions (A,B,C)  
that minimize the integral over all space of the square of  
the field difference between the molecular field and the EGF  
10 field. This integral is referred to as the **EGR Fitness  
Function (EFF)**.

$$EFF(EGR_M, F_M) = \int (EGR_M(\bar{r}) - F_M(\bar{r}))^2 d\bar{r}$$

(16)

#### Overview

20 To compare two molecules A and B, I determine the  
difference between their fields A and B using the norm of the  
field difference shown in equation (9). The norm of the  
field difference constitutes a metric distance, not just a  
similarity measure. Furthermore, the minimal value of d, dm,  
which occurs at the point of maximal overlap of the two, also  
25 forms a metric.

30 Various techniques exist to attempt to find the best  
overlap of two fields, typically involving repeated searches  
from different starting orientations of the two molecules.  
This is necessary because no direct solution for the minimal  
distance orientation is available, and most methods tend to

00644937-082300

get caught in nearby local minima, missing the global minimum. One such technique is a Gaussian technique described in J.A. Grant et al., "A Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape," J. Computational Chemistry, Vol. 17, No. 14, pp. 1653-66 (1966). Using this technique, I overlaid the two molecules shown in Fig. 2A to produce the result shown in Fig. 2B.

10 This Gaussian technique is less susceptible than other techniques to being caught in local minima. Because  $d_m$ , once found, is an invariant, fundamental distance between the two molecular fields, it can be used in creating a hyper-dimensional embedding of a set of molecules. This hyper-  
15 dimensional space is called a "shape space", (but where "shape" can mean electrostatics or other field properties, not just steric fields). I believe that for small molecules the number of dimensions will be on the order of a few dozen.

20 In the following, I may refer interchangeably to maximal overlap (or overlay), minimal field difference, and minimal distance, as they all refer to and measure the same optimal orientation of two molecules with respect to each other.

25 Rather than merely use the overlay optimization as a method of aligning molecules for further similarity comparison, I use the overlay measure as a metric distance for the organization of large numbers of structures such that searches within this dataset can be made "efficient".  
30 Efficiency here can be precisely defined in the language of database algorithms: the search time can be made "sublinear",









certain distance of a query to a constant multiplied by the  
logarithm of the number of entries in the database. For  
instance, if there are 1,000,000 entries one only has to  
check of order  $6 \times \text{constant}$  entries, and for 10,000,000 only  
5  $7 \times \text{constant}$  entries. Another method is known as "Vantage  
Trees", wherein the data is organized about certain "vantage"  
points, for instance the center of clusters of molecules in  
shape space. A third method chooses a set of key compounds,  
10 and for each of these, lists all the molecules in order of  
distance from it.

0964937-082300  
Sub. 961

For example, if I have 1000 molecules in my database I  
might organize this information thus: select 10 "key"  
molecules which are quite different in shape. For each of  
15 these 10 key molecules I then find the distance from each of  
these molecules to every other molecule in the database, and  
make 10 lists where each list has a different key molecule at  
the top and the rest of the 999 molecules are listed in order  
20 of shortest distance from it. To find the closest match  
between a test molecule and the 1000 molecules of the  
database I begin by determining the metric distances between  
the test molecule and each of the key molecules. Suppose the  
shortest distance is to key molecule 6 and that distance is  
25 X. I now begin to calculate the distances to the rest of the  
molecules, but in the order specified by that key molecule's  
list. Since the list has molecules close to key molecule 6  
first, it is likely these are also close to my test molecule.  
30 Furthermore, by the triangle inequality, since molecules  
which are a distance greater than  $2X$  from key molecule 6 must

Could  
be sub ab

be greater than X from my test molecule, I only have to go  
down the list until this condition is satisfied, i.e. I may  
not have to test all 1000 molecules. Furthermore, if I find  
a molecule closer than key molecule 6 early in the list, say  
5 distance X-d, then I only have to go down the list until the  
distance from the key molecule is greater than 2X-d, i.e. I  
can refine the cutoff distance as I progress down the list.  
Thus I can search the database, by shape, in a time sublinear  
10 with the number of molecules in the database. These methods  
are not possible without evaluating a shape space description  
of the set of molecules that comprise the database.

002280" 032300 054493" 032300

Ultimately I hope to characterize the shape space of all  
synthesizable molecules, i.e. the  $10^{200}$  referred to earlier.  
15 But the shape space of smaller sets of molecules can be  
determined as well. The "position" of a molecule in such a  
high dimensional shape space then allows one to calculate the  
best possible overlay between two molecules without any  
20 computer intensive minimization of the overlap of two fields,  
or searching through multiple minima. (The distance between  
these two points is just the square root of the sum of the  
squares of the values assigned to each coordinate in shape  
space, i.e. a higher order generalization of finding the  
25 distance between any two points in three-dimensional space.)  
I anticipate this will provide nearly a thousand-fold speedup  
in the calculation of the best overlap between two molecules,  
if their shape space positions are known, relative to my  
30 current best methods (which are already hundreds of times  
faster than any reported in the literature).



positive constants, and  $p$  may be positive or negative but is usually positive. This function falls off rapidly far from its center and has the symmetry of an ellipsoid.

5 The procedure is as follows:

*Sub a71*  
a) Chose the number of EGFs that I want to represent the field.

b) Chose random positions for the center of each EGF and make  
10 each spherical, i.e.  $a=b=c=1$ .

c) Let the center, directions and magnitudes of each EGF change such as to maximize the field overlap between this set and the molecular field. Many standard numerical techniques, e.g., steepest descent, conjugate gradient method, can  
15 perform this minimization. I have encoded the first and second derivatives of this overlap function, which leads to particularly efficient minimizations. An example of the EGF produced by this method is shown in Fig. 3.

20 The above is also a multiple minima problem, and once this procedure has been completed one typically repeats it with new starting positions, i.e. there is more than one possible representation of a molecular field by a set of EGF's. One also tries increasing numbers of EGF's, i.e. one,  
25 then two, then three, etc., as illustrated by Figs. 4A, 4B and 4C. Small drug-like molecules typically require at most 3 or 4 ellipsoids to represent them well. I can construct a measure of how many EGF's are needed by also measuring the  
30 field overlap of each EGF with the atoms that lie within its domain, where each such atom is represented as a spherical

09644937 082300  
Gaussian. This value of fit typically gets worse as the overall fit gets better (i.e. as the fit gets tighter, atoms may be split across EGF boundaries) and hence the sum of the two is a quantity which first increases with number of EGF's  
5 but then decreases, the maximal value indicating the "best" number of EGF's for that molecule, although I also keep "sub-optimal" representations since these may also be useful.

One can now use the EGF's in the molecular comparison  
10 problem in numerous ways. For instance, the simplest is to deconstruct the molecule into separate sets of atoms, each of which is then operated upon by my shape space methods, i.e. as an automatic, non-chemistry based method of fragment construction.

15 One can also use the EGF decomposition to organize the storage of molecular shape information. For instance the volumes and orientations of each EGF can be used as keys in an index for rapidly finding similarly decomposed molecules.  
20 Because the actual number of EGF's in any representation is small, e.g. 3 or 4, one can also store all subsets of EGFs. For instance a four EGF set ABCD could be stored on the basis of ABCD, ABC, ABD, ACD, BCD, AB, AC, AD, BC, BD, CD, A, B, C and D. This then provides a very simple method for subdomain  
25 matching, i.e. a molecule might have EGF's very similar to A, B and C but no fourth domain, or might have a different fourth domain from D and yet still be rapidly returned as a potential match.

30 The relationship between organizing molecular structural information by the field metric and by EGF's is synergistic.

00644937 082300

For example, rather than start with random positions for the EGF's, the algorithm will run much faster if one starts near the final solution. If one can use the overlap metric to pull up a similarly shaped molecule that has already been  
5 EGF-processed, it is very likely that the EGF composition of this molecule will form a good starting point. Similarly, if one needs shape space coordinates of a molecule relative to a set for which shape space has been characterized, then it is  
10 better to start by calculating distances to molecules that are close in shape. The EGF breakdown of the new molecule, when compared to that of those within the database, provides an "initial" embedding, i.e. by finding its rough position within the shape space.

15 In addition to representing shapes of molecules one can also use the EGF's to compare properties of the underlying atoms, i.e. to solve the "assignment problem", i.e. which atoms correspond to each other between two molecules. This  
20 is because when one compares two EGF's there are no ambiguities in their relative overlay (save for a four-fold degeneracy from rotations about the major and minor axes, i.e. there are just four ways to overlay two EGF's). As such one can either directly compare atoms belonging to each EGF  
25 (e.g. find distances from similar types of atoms to each other), or one can project such properties onto the pseudo-surface of each ellipsoid. I define a pseudo surface of an EGF as the surface of an ellipsoid that has axes in the same  
30 direction as the EGF and with the same relative axis (i.e. a/b/c) as the EGF, and which has the same volume as the EGF.







## IMPLEMENTATION AND EXAMPLES

Specific implementations and examples of the foregoing include the following:

- 5        1: Finding The Maximal Overlap (Minimal Field Difference) Between Two Fields A And B
- 2: Refining The Search Position Via Numerical Or Analytical Derivatives
- 3: Determining The Shape Space Of A Set Of Molecules
- 4: Calculating The Position Of A New Structure In A Preconstructed Shape Space
- 5: Extending The Shape Space
- 6: Calculating The Maximal Overlap Between A New Structure And A Large, Previously Shape-Space Decomposed Set Of Molecules
- 10       7: Using The Shape Space Description To Correlate With Known Biological Activity
- 8: Examples Of Using The Minimum Field Difference Metric To Organize A Database Of Molecules
- 9: Examples Of Using The Shape Space Positions To Organize A Database Of Molecules
- 10: Local Domain Decomposition
- 15       11: Constructing An Ellipsoidal Gaussian Representation (EGR).
- 12: Construction Of Multiple EGR's Containing The Same Number Of EGF's.
- 13: Constructing Molecular Fragments From An EGR.
- 14: Evaluating An EGR Fit: The Fragment Adjusted EFF.
- 15: Construction Of Multiple EGR's With Different Numbers Of EGF's
- 20       16: Storage In A Database
- 17: Comparing Single EGR's To Solve The Atom Assignment Problem
- 18: EGF Pseudo-Surfaces And Painted EGF's
- 25       19: Using A Database Of EGR's For (Molecular/Molecular Fragment) Similarity Evaluation
- 20: Using EGF's In Similarity Searches
- 21: Using EGR's To Find A Place In Shape Space; Using Shape Space To Find EGR's
- 30       22: Defining An EGR For A Negative Space

- Sub A85
- 1: Finding the maximal overlap (minimal field difference) between two fields A and B difference) between two fields A and B

Exhaustive search.

5 The optimal overlay of two molecules depends upon six variables, i.e. three translational degrees of freedom and three rotational degrees of freedom. Since the time taken for any search technique rises as the power of the

10 dimensionality, the problem of optimal overlay is computationally expensive. However, the metric property of the field difference allows for certain improvements in performance because of the triangle inequality. Thus

15 suppose I am trying to overlay molecule B on molecule A and I already have one relatively "good" overlay, with a field difference value of d. Suppose I have just calculated the field difference value for a particular orientation and central position of B with respect to A, d', which is

20 greater than d. I am now in a position to determine which orientations and central positionings relative to this orientation and positioning might have a field difference of less than d. To see this note that the triangle

25 equality lower bound on metric properties states that:

$$|d(AB) - d(BC)| \leq d(AC)$$

where d(AB) is the field difference between molecule A and molecule B at a given orientation and translation relative to each other, and similarly for d(BC) and d(AC), and the

30



set of orientations and separations will range over all such that give an overlap greater than some critical value close to zero.

- 5 (ii) Find what might be a good overlay between A and B so as to establish a reasonable value of d. One does this by trying a set of translations and orientations, B'', to coarsely sample all reasonable overlays (as in (i), this means that the overlap is greater than some small threshold value), and setting  $d_{\min}$  equal to the best such value obtained.
- 10 (iii) Sample around each such overlay tried in (ii) using translations and rotations B''' of B'' from these positions as precalculated in (i), such that the triangle lower bound can be used to prevent needless calculation of possible overlays.
- 15 (iv) Accept the best overlay from (iii) as the best possible overlay, or select the best N such as candidates for further refinement, e.g. via numerical optimization as described below.
- 20

#### B) Selective Search:

25 Align the two fields based upon aspects of the structures or fields that suggest this alignment might be good. For instance, if the two molecules are long and thin then it makes sense to align them such that these axes are in the same direction. Other examples might be to align similar chemical fragments. In my approach I calculate quantities

30



set. Note than N will not be the total set of molecules and may be as small as two.

(b) Given that the norm of the minimal field difference

5 between any two molecules is a metric distance,  
construct the distance matrix D, where the element ij of  
D is the minimal field difference between molecule i and  
molecule j.

(c) Construct what is known as the metric matrix G from D, as  
10 described in any description of distance geometry, e.g.  
Blaney and Dixon, "Distance Geometry in Molecular  
Modeling", Reviews in Computational Chemistry, Volume V,  
VCH publishers, 1994.

(d) Diagonalize G using any standard technique to find the  
15 eigenvalues of a matrix.

(e) From this diagonalization procedure find a set of  
positions in N-1 dimensional space that reproduce the  
distances in the matrix D (see the Blaney/Dixon  
20 reference for details)

(f) Determine which coordinates can be set to zero for each  
and every molecule and still enable the distance matrix  
to be reproduced to within a given tolerance T. (This is  
equivalent to setting all eigenvalues of absolute  
25 magnitude less than some cut-off value to zero). The  
procedure for this is to a) set all the coordinates I  
want to keep to zero, b) find the geometric center  
(average coordinates) of the remaining coordinates (the  
ones I might be able to discard), c) find the largest  
30 distance from this center to any one point. d) ascertain





- 00330" 22644960
- (ii) The largest void within the space, i.e. the largest hypersphere of the same dimensionality as the shape space that can fit between molecules. This can be used to ascertain which molecules from another set would best fill that gap and hence make the coverage of shape space more uniform.
- (iii) The volume of the space occupied by the complete set of molecules (i.e. the volume of what is known geometrically as the "convex hull" defined by those points in the shape space). This quantity is useful in the context of deciding what fraction of the shape space a subset of molecules covers compared to the complete set.
- (iv) The smallest subset of molecules which reproduces (iii), i.e. molecules whose shape space positions lie on the convex hull of that set of molecules. These molecules define the boundaries of shape space and hence are useful as the smallest subset of molecules which has the same shape space volume as the total collection.
- (v) The local dimensionality around a particular molecule. Given a distance cut-off and a particular molecule, I can calculate the dimensionality of the local shape space of the set of molecules consisting of this molecule and all those closer than the cut-off. The use of this is that certain subsets of molecules may embed within the global shape space in a space of much lower dimensionality (imagine a set of points lying on a curved surface in 3 dimensional space; the global

dimension is 3 but the local dimension is 2). This has import for the efficient storage of the shape information of molecules.

4: Calculating the position of a new structure in a preconstructed shape space

5  
10  
15  
20  
25  
30

Once I have a shape space for N molecules, of dimension M, the next step is to calculate the position within this shape space for a molecule not used in the construction of that shape space. This position is found by analogy with triangulation in three dimensions, i.e. if one has a set of distances from an object to four reference objects the exact position can be ascertained. In two dimensions one needs three distances. In M dimensional shape space one needs M+1 distances. (In each of these cases, the M+1 distances must be from points which cannot as a set be described at a dimensionality less than M, i.e. for the case of three dimensions, the four reference points cannot all lie in a 2 dimensional plane). The actual procedure for going from distances to a position is simply that a linear equation for the coordinates can be generated from each distance, such that the solution of the set of such produces the position. This set of linear equations can be solved by any standard method, for instance, Gauss-Jordan elimination (see, for example Stoer and Bulirsch, "Introduction to Numerical Analysis", 2<sup>nd</sup> Ed., Springer-Verlag, chapter 4). An important note here is that this procedure can fail, i.e. it will produce a position which will underestimate the M+1 distances by a constant amount. This is an indication that the

structure under study actually lies in a higher dimensional space than the shape space previously constructed. As such, that shape space needs to be extended.

5: Extending the Shape Space

5

If the position determining procedure fails, then it is necessary to increase the dimensionality of the shape space description. This is straightforward to accomplish in that I merely need to add an additional variable to all the previous

10 positions, set the additional variable to zero for these structures, and to then find what the new M+1 coordinate for the new molecule needs to be such that the new coordinates now reproduce the distances correctly. But this is simple to calculate from the shortfall in distances calculated by the  
15 previous method, since there is only the need to add one more coordinate.

6: Calculating the maximal overlap between a new structure and a large, previously shape-space decomposed set of molecules

20

Suppose I have a set of N structures for which a shape space decomposition of dimension M is known. If I have a new structure and wish to find the closest such structure within this set I could use the minimal field distance method  
25 between this one molecule and all molecules within the set.

Many of the ideas within this document concern how to **avoid** doing this since N may be very large. An additional component is avoiding the performance of more applications of the minimal field distance method than are necessary.

30



predict the activity of molecules for which the input quantities are known, but not the biological activity. The use with the shape space decomposition is as follows:

- 5 (i) Calculate the shape vector for each molecule for which a biological activity value is known. Note here that the shape vector can be relative to a shape space defined by a completely separate set of compounds, or to the space  
10 calculated from that very set of compounds.
- (ii) Use the numbers that make up this vector as input to a PLS procedure, with or without other quantities known for each molecule under consideration.
- 15 (iii) Use the resultant "weights" to predict the activity of other molecules not in the original set, i.e. by calculating their shape space vector.

Note that more than one shape field can be used as input to a  
20 PLS calculation, i.e. the shape vector for the electrostatic  
field as well as that for the steric field.

8: Examples of using the minimum field difference metric to organize a database of molecules

Required is a set of N molecular structures. (These  
25 will belong to L molecules where L can be less than N if  
there is more than one conformer of a molecule in the set.)  
These structures may also have unique chemical identifiers  
(e.g. chemical names, SMILES strings, catalog numbers etc).

30 1) Constructing and using a Distance Tree

July  
Pao

(i) Chose a structure at random from the N possible structures.

5 (ii) Find the field distances from this root structure to all N-1 other structures.

(iii) Calculate the median of the distances found in (ii).

(iv) Use the median value as a threshold distance T to subdivide the N-1 other structures into two lists, or  
10 "branches", based upon this criterion, with the lower branch containing all structures with distances below the threshold, and the upper branch containing all structures with distances greater than the threshold.

15 (v) Store the threshold value along with the root structure in the root node data structure.

(vi) Repeat this process for each list from i) onwards, but with N decremented by one until the repeatedly divided trees are of size one or zero

20

Now, faced with a problem of finding the closest structure in the database to a novel example, i.e. one not in the database, I proceed as follows.

(i) Find the distance to the root structure.

25

(ii) If this distance is less than half the threshold distance ( $T/2$ ) for this node, then I need never check any structure along the upper branch of the tree, which contains structures whose distance from the root is greater than that  
30 threshold, since by the lower bound of the triangle



002280" 2654960

(iii) For each of the  $K$  key structures, find the minimal distance  $m$  from it to every other structure in the database. If  $K$  is large, this step may also be speeded  
5 if the shape space has been determined, allowing simple distance calculations rather than complete overlay calculations for every structure.

(iv) For each of the  $K$  key structures, create a list associated with it, and place into the list, in order of  
10 increasing distance, references (name, number etc) to each database structure along with its distance  $m$  from the key structure.

15 B. Using the lists to find the structure closest to a test molecule.

*Sub B3* As illustrated in the flow chart of Fig. 7, the closest structure is found by the following steps:

20 (i) For a test molecule, find its minimal distance  $x$  to each of the  $K$  key molecules.

(ii) Choose the list whose key molecule  $k$  is closest to the test molecule, where this distance is  $X$ . Since the list  
25 has molecules close to  $k$  first it is likely these are also close to my test molecule.

(iii) Set as current list position  $n$  the top of list  $k$ .  
Set a variable *BEST* equal to  $X$ .

30



(iv) Otherwise, if the distance  $m$  from key  $k$  to list structure  $n$  is greater than  $X + BEST$ , then stop, as by the lower bound of the triangle inequality no structure further  
5 down the list can be closer to the test molecule than *STRUCT*, i.e. if  $m = X + BEST + a$  for any positive distance  $a$ , then the triangle inequality  $|m - X| < d$  can be rewritten as  $d > BEST + a$ .

10 (v) Find the minimal distance  $d$  from the test molecule to the current structure  $n$  on the list.

(vi) If  $d < BEST$ , store  $d$  in *BEST*, and  $n$  in *STRUCT*.

(vii) If more structures, increment the list position  $n$   
15 by one and continue at (iv).

(viii) When the procedure terminates, the index  $n$  of the closest structure to the test molecule is found in *STRUCT*.

20

*data 12*  
Thus I can search the database, by minimum field difference, in a time sublinear with the number of molecules in the database. This is because, by the triangle inequality, I know the cutoff distance for evaluating structures in the  
25 list is at most equal to  $2X$  (when  $BEST = X$ ) and is potentially further refined as I progress down the list and find better (smaller) values for *BEST*. As noted above, the list creation process can be speeded if the shape space of  
30 the structures has already been determined. Whether the time

Cont'd  
Jul. 1962  
912

saved will be justify the time spent constructing the shape space depends on the number of key structures  $K$  and the number of structures in the database.

5 9: Examples Of Using The Shape Space Positions To Organize A Database Of Molecules

Making and Using an M-dimensional tree:

M dimensional data points may be stored in a tree-like data structure such that an efficient search can be made to find all points within a distance  $d$  of a new point. These algorithms are standard in art. Although the performance of this tree lookup is not guaranteed efficient, i.e. there are pathological cases where it is no better than testing all points, on the average it allows the number of search steps to be reduced from  $N$  (the number of points in the tree) to some multiple of the logarithm of  $N$ .

15 1) Constructing and using an M-dimensional tree.

(i) Find the shape space positions for a set of  $N$  structures.

Jul. 1962  
a320 (ii) Chose a structure at random from this set and record its name in the zero level node of a tree structure which is such that each "node", or "slot", has two child nodes, called "left" and "right", at what I refer to as a level one greater than this node.

25 (iii) Select a second structure at random and store its name in either the left level one node of the tree if its first shape space coordinate is less than that of the first structure, otherwise place it in the right level one node.

30 (iv) Chose another structure at random. As before, test its first shape space coordinate. This time, however, if the



(vii) Perform the same test as in (ii)-(v) on each of these structures, except that the test is now performed using the second coordinate.

5 (viii) Proceed down the tree, adding the "hit" list where appropriate, culling portions of the tree were possible until all structures have either been tested for addition to the list or culled.

*Jul. 21/4*  
10 In (1) above, rather than choosing structures at random for insertion into the tree, they could instead be sorted into a list, for example in order of increasing volume, and then taken sequentially from the list for insertion into the tree. This allows additional criteria to be used to terminate a  
15 search of the branches of the tree.

#### 10: Local Domain Decomposition

The overlay of molecular fields is a method of finding global similarities between molecules, i.e. whether molecule  
20 A has the same distribution of properties as molecule B. While this is extremely useful, it is also the case that local similarities are of interest, i.e. when a part of molecule A is similar to a part or all of molecule B. The  
25 methods described below are an approach to solve this problem by defining representations of parts of a molecular field which have an approximately ellipsoidal character. These representations conform to visual and chemical intuition as to possible fragmentations of a molecular structure, and can

30



(i) Define an analytic description of the molecular field. An example would be the steric field generated by representing each atom by a Gaussian function (defined above) and forming the sum or exclusion product field (defined above) from each such Gaussian.

(ii) Given such an analytic description the EFF becomes the sum of a series of integrals since the square of the difference field is now a sum of integrable functions. If the molecular field is defined from Gaussian functions, then the integrals are of a particularly simple form.

15 Given a procedure to calculate the EFF, the EGF parameters can be optimized to minimize the EFF. Standard methods exist in the literature for such an optimization, whether the EFF is calculated numerically or analytically.

Of significance to the practical implementation of any  
20 method are the initial EGF parameters. My work suggests that  
the initial parameters must be such that there is significant  
overlap between the each EGF and the molecular field. Thus,  
my typical starting configuration is a center for each EGF  
within the molecule (defined as within an atoms radius of at  
25 least one atom within the molecule), with axes A, B and C set  
to the x, y and z Cartesian axes, and widths u, v and w set  
to 1.0, and prefactor p set to 1.0.

12: Construction Of Multiple EGR's Containing The Same Number Of EGF's.

00644937 082300  
006280 2644960

Most methods of optimizing EGF parameters, such as minimizing the EFF, suffer from one drawback, namely that they converge to a set of parameters that do not necessarily correspond to the lowest possible EFF value, rather such  
5 parameters are "locally stable", i.e. any small change in any one parameter results in an increase in the EFF value. Such "multiple minima" are intrinsic in the nature of the problem. The local minimum with the lowest EFF value is referred to as  
10 the "global minimum". I utilize the multiple minima nature of the problem in that I do not necessarily require a single EGF for a particular molecular structure, in fact I welcome multiple representations for uses described later in this document.

15

Procedure:

- (i) Choose how many EGF's I want in the EGR.
- 20 (ii) Find the EGR by optimizing the EGF parameters to minimize the EFF.
- (iii) Add the characteristics of this EGR (i.e. the EGF parameters) to a list of possible EGR's.
- (iv) Find a new EGR by repeating (ii) but with different  
25 starting positions (which, as described above, are chosen to have random centers within the molecule).
- (v) Compare this EGR's characteristic to all those on the EGR list.
- 30 (vi) If these characteristics are "similar" to any on the list discard this EGR, otherwise add it to the list.





representation is the more appropriate representation, i.e. lower EFF means a better fit. Comparing EGR's with different numbers of EGF's is more problematic because the number of adjustable parameters is proportional to the number of EGF's, hence the more EGF's the lower the EFF tends to be. Yet, more EGF's in an EGR are not necessarily a more useful domain decomposition since this will ultimately tend towards one EGF per atom. Decomposing a molecule into each of its component atoms is a trivial exercise of little utility. A more useful measure of fit can be obtained as follows:

- (i) Calculate the molecular fragments induced by the EGR as described above.
- (ii) Calculate the EFF of the field produced by each  
15 molecular fragment with the EGR to which it has been assigned.
- (iii) Sum these domain EFF's together to form a quantity  
I will refer to as the **sum fragment EFF**.
- 20 (iv) Calculate the EFF of the whole molecule's field vs. its EGR. I will term this quantity the **molecular EFF**.
- (v) Add together the sum fragment EFF and the molecular EFF to form what I refer to as the **fragment adjusted EFF**.

25

Although the fragment adjusted EFF typically is smaller for two EGF's versus one, I have observed that as the number of EGF's further increases, the fragment-adjusted EFF eventually starts increasing again. The number of EGF's for which the





from X with the largest value from its u, v, and w values is aligned with the axis from Y with the largest of its u, v and w values and similarly for the smallest of such coefficients. There are actually four different ways this can be achieved, given the symmetry of an ellipsoidal function.

5

July 9/16  
10 (iv) For each of the four alignments, make the atom to atom assignments for the atoms which belong to the pair of EGF's being aligned together based upon "closest" or "closest of similar type".

15 (v) Rather than have an infinite number of possible alignments I now have just four to chose from, and given any kind of measure for the assignment (e.g. minimize the sum of the distances of each atom pair) this is straightforward.

Note that if the number of EGF's for X and Y is one then this provides a match between all atoms of X with all of those of Y. Otherwise, the choice of EGF's from X and Y induce a partial match between X and Y. Given a measure of the quality of this partial assignment, all possible combinations of EGF's between X and Y can be chosen to find the best EGF induced assignment.

25

#### 18: EGF Pseudo-Surfaces And Painted EGF's

Rather than compare molecular fragments via an atom assignment procedure one can compare the properties in the vicinity of each fragment. To do so I introduce the concept

30



The use of the EGF pseudo-surface lies in the fact that if properties are assigned to each point on this surface, then such properties can be compared with those belonging to a different EGF pseudo-surface. I refer to an EGF with a pseudo-surface that has properties associated with it as a **"painted EGF"**. Properties can "paint" an EGF in a variety of ways, for example:

- (i) Be analytically defined on the surface of the ellipsoid from the underlying analytical form of the property.
- (ii) Be calculated at a number of "sample" points distributed on the surface of the ellipsoid, either from analytic functional forms of the property, or interpolated from values of the property at nearby points in space (e.g. from a grid of values).
- (iii) Be assigned values at sample points based upon proximity to underlying atoms, e.g. the atomic properties are "projected" to this surface.

The information in the pseudo-surface values at a set of points (as in (ii) above) may be stored in several ways. (a) The first is just as a list of values associated with each point. (b) If the values at points are of binary nature (one or zero) then the storage may take the form of a bit-pattern. (c) The pattern of values at points may be transformed into an approximate functional form, typically spherical harmonics, i.e. an analytic form as (i) above. (d) Finally, the patterns may be stored "virtually", i.e. not actually

calculated until required but rather "stored" as a method of calculation.

5 The comparison of the properties "painted" on two such surfaces is as follows:

- 10 (i) Align the two EGF's relative to each other in one of the four possible orientations wherein centers and axes are aligned as above.
- (ii) Change the shape of each pseudo-ellipsoid to that of a sphere of unit radius via a series of three compressions, one along each axis of the ellipsoid.
- 15 (iii) Compare the properties, point by point, of the now identically shaped surfaces and superimposed surfaces.

002230 " 082300 09644937

The comparison step in (iii) can actually be made a  
20 metric distance, i.e. if the integral over the difference in properties at the surface of the unit spheres is found, then the square root of this quantity is a metric distance for exactly the same reason the similar quantity is such for three dimensional fields. Technically speaking the field  
25 described by a property on the surface of a unit sphere has the mathematical designation  $S^2$ , as opposed to  $R^3$  for a molecular field, but is still a field. However, the problem of finding the minimal distance between two such spherical  
30 fields is much simpler in my case than for general  $S^2$  fields because the overlay of the ellipsoids only allows for four









expected between the new structure and the current structure from the database.

(viii) If this estimate of the best EFF between the new structure and the current database structure is greater than the BEST parameter go to (v).

(ix) Otherwise actually find the best metric field difference between the new molecule and the current database structure. If this value is less than BEST, set BEST equal to this value, set the value of BESTSTRUCTRE to indicate this structure. Go to (v) unless this is the last structure in the database.

In the above procedure I am using the EGF's that make up the best EGR of a molecule to save time in actually having to calculate the best orientations between two molecules, which is analogous to the use of the triangle inequality of the metric field difference distance. In fact I am using this metric nature, but in a more convoluted way than before.

Note that I can improve on the above method by organizing the structures based upon the EGF's that make up the best EGR for each structure, i.e. such that instead of starting with the first structure in the database I start with the structure which has the most similar set of EGF's ("most similar" defined here as having the most similar volumes). Starting with a structure that is more likely to have a small resultant BEST parameter means that step (vii) is more likely to reject the next structure, i.e. the time intensive step of finding the best overlay is avoided more



reproduced then finish the procedure, otherwise go to  
(iii).

5 The rationale behind this procedure is that although the  
shape space dimensionality of the database of structures may  
be  $M$ , the local dimensionality of molecules like the new  
molecule may be much less than  $M$ . As an example, if in three  
dimensions points lie on a line which is curving through  
10 space, and if my test point lies on this line, close to two  
other points, then the distances to just these two points can  
define quite accurately the position on this line, not the  
normal four such points that would be required if the points  
were randomly distributed in all three dimensions. Hence,  
15 given similar structures I find the position in shape space  
with fewer than  $M+1$  comparisons, if the local distribution of  
such structures is not smooth in all  $M$  directions.

Given a shape space vector of a new molecule but not an

20 EGR

- (i) Find the most similar structure, or set of structures  
from within a database of such.
- (ii) Use the EGF positions of each of the EGR's associated  
with each such structure as a starting point for an EFF  
25 minimization for the new molecule.
- (iii) Then perform a "normal" EFF minimization but with  
random starting positions.

30 The rationale here is that since the structures found in  
the database are similar then the EGR's will also be similar.

If so, much time will be saved in step (iii) because many of the potential EGR's will have already been found.

22: Defining An EGR For A Negative Space

5 Of particular interest in the use of EGR descriptions is finding molecules that might fit in the active site of a protein, as this is the mode of action of most pharmaceutical compounds. The procedure is as follows:

10 Finding active site EGF's:

*Handwritten: A18*  
15 (i) Define a fitting function  $f$  between any two EGF's such that if both were spherical this function would be a minimum when the inter-EGF distance is the same as the sum of the radii of each EGF (defining the radii of the EGF as that of a sphere of equivalent volume). Such a function for two EGF's, EGF1 and EGF2, is:

20 
$$f = a*V - b*(Q(EGF1, V) - b*(Q(EGF2, V)$$

where  $V = Q(EGF1, EGF2)$  where  $Q$  is defined in equation (6) above.

25 (ii) Define a molecular field function from either a sum or exclusion product of a set of  $N$  atomically centered spherical EGF's, where each such EGF has the same volume as the atom upon which it is placed, and where the  $N$  atom centers belong to all of the atoms in the protein  
30 within a specified distance of the active site (the active site may be defined in various ways, for instance



